TILBURG UNIVERSITY

Tilburg University

Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results)

Wicherts, J.M.

Published in:
Intelligence

DOI:
10.1016/j.intell.2016.11.002

Publication date:
2017

Document Version
Peer reviewed version

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Wicherts, J. M. (2017). Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). *Intelligence*, *60*, 26–38. https://doi.org/10.1016/j.intell.2016.11.002

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 04. May. 2023

Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results)

Jelte M. Wicherts

Tilburg University

In press, Intelligence, November 2016

Author note: The preparation of this article was supported by a VIDI grant (no. 452-11-004) from the Netherlands Organization for Scientific Research.

Abstract

Spearman's hypothesis stating that ethnic group differences on cognitive tests are most pronounced on the most highly *g* loaded tests has been commonly tested with Jensen's method of correlated vectors (MCV). This paper illustrates and explains why MCV applied to item-level data does not provide a test of measurement invariance and fails to provide accurate information about the role of *g* in group differences in test scores. I focus on studies that applied MCV to study group differences on items of Raven's Standard Progressive Matrices (SPM). In an empirical illustration of the psychometric problems with this method, I show that MCV applied to 60 SPM items incorrectly yields support for Spearman's hypothesis (so-called Jensen Effects suggesting that the group difference is on g) even when the items in the second group are not from the SPM but rather from a test composed of 60 items measuring either anxiety and anger or the big five personality traits. This shows that MCV applied to item level data does not accurately reflect the degree to which item bias or *g* play a role in group differences. I conclude that MCV applied to items lacks both sensitivity and specificity.

Keywords: group differences, race differences, psychometrics, differential item functioning, measurement invariance

1. Introduction

Spearman's hypothesis states that ethnic group differences on cognitive tests are due to *g* (Jensen, 1985), and hence that observed ethnic group differences on these tests cannot be attributed to lower-order cognitive ability factors or measurement bias at the test or item level. Twelve recent studies used the method of correlated vectors (Jensen, 1998) to test Spearman's hypothesis with scores of different ethnic groups on various versions of Raven's Progressive Matrices (Díaz, Sellami, Infanzón, Lanzón, & Lynn, 2012; Rushton, 2002; Rushton, Bons, Vernon, & Cvorovic, 2007; Rushton, Cvorovic, & Bons, 2007; Rushton & Skuy, 2000; Rushton, Skuy, & Bons, 2004; Rushton, Skuy, & Fridjhon, 2002, 2003; te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al., 2015; te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev, & Repko, 2015; te Nijenhuis, Bakhiet, et al., 2016; te Nijenhuis, Grigoriev, & van den Hoek, 2016). The goal of these studies was to test whether ethnic group differences were most pronounced on Raven's items that showed the highest loading on *g.* To this end, vectors of ethnic group differences on Raven's items were correlated with vectors representing the degree to which these Raven's items correlated with the *g* factor. Significant correlations from this method of correlated vectors are called Jensen Effects (Rushton, 1998). Jensen Effects are seen as supporting Spearman's hypothesis and are taken to mean that ethnic differences are "not explainable in terms of test bias or in terms of differences in types of item content or other formal or superficial characteristics of the tests" (te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al., 2015, p.119). Jensen Effects are accorded a central role in the debate on the nature and nurture of ethnic group differences in cognitive ability test performance (Jensen, 1998; Rushton, 2002; Rushton & Jensen, 2005), and are often invoked as evidence in favor of a genetic component to ethnic differences (Rushton, Bons, et al., 2007). Moreover, finding a Jensen Effect is considered relevant for use of the

tests in practice because it appears to suggest that the test at hand can be safely used to make inferences about test-takers' latent ability regardless of their ethnic background.

A substantial literature addressed the drawbacks of the method of correlated vectors (Ashton & Lee, 2005; Dolan, 1997, 2000; Dolan & Hamaker, 2001; Dolan & Lubke, 2001; Dolan, Roorda, & Wicherts, 2004; Lubke, Dolan, & Kelderman, 2001; Millsap, 1997; Wicherts & Dolan, 2010; Wicherts & Johnson, 2009), but the method continuous to be used commonly. The goal of this paper is to discuss in non-technical terms the method of correlated vectors (MCV) to study Spearman's hypothesis at the item level. MCV applied to items revolves around item-total correlations as measures of items' loadings on the *g* factor, and the group difference in proportions correct on each item, or, in other words, the group differences in items' p-values. I will discuss drawbacks of the use of such *classical test theory (CTT)* item statistics that have been known since the 1940s (Ferguson, 1941; Gulliksen, 1950), and inspired the development of modern item response theory or IRT (Embretson & Reise, 2000; Lord, 1980; Lord & Novick, 1968). A fundamental difference between CTT and IRT is that in the former framework item statistics are operationalized on the basis of observed item scores (here: correct or incorrect) while in IRT the items parameters are defined vis-à-vis the latent ability that the test purports to measure. One crucial implication is that CTT item statistics (like p-values and item-total correlations) are necessarily different between groups that differ in latent ability (Embretson & Reise, 2000), whereas IRT item parameters can be meaningfully compared across groups. IRT allows a rigorous test of whether the items in a scale function equivalently across different groups (i.e., display no Differential Item Functioning or DIF), which is a crucial requirement for any meaningful interpretation of group differences in terms of latent variables such as *g*. Because CTT does not offer tests of measurement invariance that involve latent

variables, CTT-based methods (such as MCV) are ill equipped to study whether group differences on item performance can be attributed to the targeted latent variable(s) or to measurement bias. Another problem with CTT is that it focuses on the "true score", which cannot be equated with the construct that the test is supposed to measure (Borsboom & Mellenbergh, 2002). Even nonsensical tests composed of heterogeneous items tapping on widely different constructs have a true score as defined in CTT, as I will illustrate below by adding items from different mood and personality scales. Because MCV at the item level uses this true score as means to operationalize the targeted trait (here *g*) and the degree to which items correlate with that targeted trait (here the *g* loading), MCV could lead to incorrect assessments of the role of *g* in group differences on the items when in fact the true score does not accurately reflect *g*.

In this article, I will first introduce MCV by focusing on how it was originally developed (Jensen, 1980, 1998), namely for studying group differences on subtests from a larger cognitive ability (IQ) test battery with linear factor models. Subsequently, I will discuss four problems with MCV applied to item level data (see also: Wicherts & Johnson, 2009), concerning its inability to test measurement invariance, the group-specificity of item-total correlations, the unwarranted interpretation of item-total correlations as *g* loadings, and the complex non-linear relations between the vectors in MCV. Finally, I present the results of an empirical study of what happens with MCV if we replace cognitive test items with items from entirely unrelated scales measuring anger, anxiety, and personality in one of the two groups that are being compared. These results are valuable in assessing whether MCV is capable of detecting instances in which item bias and DIF can hardly be any more severe simply because items measure different traits across groups.

2. Method of correlated vectors with subtests

Spearman's hypothesis states that ethnic group differences are due to *g*, implying that the degree to which any cognitive subtest shows group differences can be predicted by the degree to which each subtest measures *g*. In its original form, MCV (Jensen, 1980, 1998) uses *g* loadings based on a factor analysis of the subtests within the two groups that are being compared. Subsequently, these *g* loadings are put in a vector that is as long as the number of subtests. Next, the between-group mean differences on each subtest are computed, and some effect size measure (typically Cohen's *d*) will indicate how strongly the two groups differ on each of these subtests. The crucial test of Spearman's hypothesis in MCV is the correlation between the vector of subtests' *g* loadings and the vector of group differences on the same subtests. Significant MCV correlations (tested against a correlation of 0 using as N the number of subtests) are then called Jensen Effects (Rushton, 1998, 2002).

Jensen (1998) reported that the typical MCV correlation based on cognitive subtests and applied to Black-White differences in the United States is around .63. Since that time, a great deal of research addressed the factor analytic technicalities of MCV applied to subtests (Ashton & Lee, 2005; Dolan, 1997, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001). The main conclusion from these studies is that MCV applied to subtests might lead to misleading results, or as te Nijenhuis (2013) called it: "The method of correlated vectors is not a strong statistic [sic]" (p. 228).[1]

Several empirical studies (Dolan & Hamaker, 2001; Dolan et al., 2004) used multi-group confirmatory factor analysis (MGCFA) and found that large Jensen Effects with MCV can occur even if *g* is *not* the main (or only) source of the group difference as

[1] te Nijenhuis (2013) proposed to combine MCV with psychometric meta-analytic approaches (Hunter & Schmidt, 2004). It is beyond the scope of this paper to discuss these extensions, which have been applied in a number of papers (e.g., te Nijenhuis, Jongeneel-Grimen, & Kirkegaard, 2014), but whose technical specifics have not been studied formally.

evidenced by substantial violations of measurement invariance and group differences in first-order factors. These results are problematic because the study of Jensen Effects aims at distinguishing between two alternative hypotheses, one in which *g* explains the group difference (Spearman's hypothesis) and another in which other factors (item bias, subtest-specific abilities, or other non-*g* factors) play a role in the observed group differences in test or item performance. In terms of diagnostics, high sensitivity would imply that if *g* is indeed the only source of the group difference, the MCV correlation should be close to 1. On the other hand, if *g* is not the (only) source of the group difference, the MCV correlation should be close to zero (or perhaps even negative), thereby supporting MCV's specificity. Dolan and colleagues (Dolan & Hamaker, 2001; Dolan et al., 2004) showed both empirically and formally that MCV applied to the subtest level exhibits weak specificity because Jensen Effects can occur even if *g* is not the main source of group differences (a false positive in diagnostic terms). Ashton and Lee (2005) studied scenarios wherein Spearman's hypothesis was true while MCV (at the subtest level) yielded low correlations. In terms of diagnostics, this means that false negatives are likely and hence that MCV applied to the subtest level data has weak sensitivity (which does not mean that true positives or true negatives cannot also occur in MCV; Dolan, 1997; Dolan & Lubke, 2001).

3. MCV does not yield a test of measurement invariance

A comparison of cognitive test scores across groups in terms of latent variables requires that the tests or items are measurement invariant with respect to these groups. Measurement invariance is a core requirement for Spearman's hypothesis stating that groups only differ in the latent variable *g* (Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001; Lubke, Dolan, Kelderman, & Mellenbergh, 2003a, 2003b).

Mellenbergh's (1989) general definition of measurement invariance focuses on

the distribution (in his formulation expressed with P) of observed test or item scores *X*, conditional on the latent variable *theta* that the test or item is supposed to measure, and a group indicator *v*. Measurement invariance with respect to groups based on *v* holds if:

$$P(X \mid \textit{theta}, v) = P(X \mid \textit{theta}). \quad (1)$$

This definition uses conditional distributions (indicated by "P(|)") that describe the distribution of scores on X after we have taken into account the scores on the latent cognitive ability theta within the groups. Specifically, the definition states that the distribution of observed scores X, which is conditional on the latent cognitive ability (theta), does not also depend on the grouping variable *v*. This definition is general as it underlies both tests of measurement invariance in the linear factor model (Meredith, 1993) and tests of measurement invariance at the item level (Holland & Wainer, 1993). When considering items, P in Equation 1 denotes the chance of answering the item correctly, conditional on the targeted trait (*theta*) and the group indicator *v*. If we replace *theta* with *g*, the definition of invariance offers another way of expressing Spearman's hypothesis for dichotomous items in a scenario where the test measures only *g*. In this hypothetical case, invariance implies that two individuals with the same level of *g* should have the same probability of answering X correctly, regardless of the (ethnic) group *v* they are in. This would mean that observed group differences in item performance are due only to *g*, as hypothesized by Spearman. If the equality in (1) does not hold, this suggests that there are additional sources of group differences on X above and beyond *theta (g* here*)* that violate Spearman's hypothesis.

Under measurement invariance group differences in observed scores X can be safely attributed to group differences in *the latent ability that the test measures*. Because cognitive tests measure additional factors beyond *g* (Spearman denoted these test-specific abilities by *s*), finding measurement invariance with a given cognitive scale is a

necessary but not a sufficient condition for Spearman's hypothesis (see also Section 5). Even finding measurement invariance with respect to a latent first-order factor in a battery of subtests is insufficient for claiming support for Spearman's hypothesis. For instance, if three indicators of verbal intelligence exhibit measurement invariance in a factor model, we still do not know whether the group difference in this latent verbal ability factor is on *g*. This problem is even more pronounced when one uses a single subtest like the Raven's test. To test Spearman's hypothesis, one would need more indicators of *g* besides verbal intelligence (e.g., working memory, spatial ability), test these additional indicators for invariance also, and verify that the group differences is attributable to the higher-order *g* factor and not verbal intelligence (or any of the other first-order factors) in particular.

There exist several ways to test measurement invariance at the item level. Earlier methods like the Mantel-Haenszel procedure or logistic regression used observed sum scores on the scale at hand for conditioning, whereas more refined latent variable methods involve conditioning on the latent variable of interest (Millsap & Everson, 1993). MCV at the item level has some resemblances with logistic regression approaches to testing DIF, wherein one regresses dichotomous item performance on both the total sum score on the test and the group indicator (and perhaps also an interaction between group and the total sum score). The main differences between MCV and logistic regression lie in the incorrect use of a linear model in MCV (relations with dichotomous item scores are nonlinear; see Section 6) and the lack of consideration of individual items in MCV. In logistic regression tests of DIF, the regression is appropriately logistic and the focus is on studying whether group differences on the item remain after controlling for group differences on the sum score. Such conditioning is absent from MCV, hence it does not entail a test for DIF or measurement invariance.

Latent variable procedures to testing DIF are well established (Holland & Wainer, 1993) and involve the test of equality across groups of the item parameters defined within a particular item response model, such as the logistic Rasch model. Group differences in item parameters across groups (i.e., DIF) imply that additional factors besides the latent variable of interest are responsible for the group difference observed on that item. Whereas IRT item *parameters* should be identical across groups under invariance, the same does not apply to item *statistics* from CTT, like the p-value or item-total correlations. Particularly relevant for MCV is that item-total correlations are normally different across groups differing in latent ability even when measurement invariance holds.

4. Item-total correlations differ for groups differing in ability

In applications of MCV at the item level, correlations between the dichotomous item score and the total sum score on the test, or item-total correlations, are considered to be the item's *g* loading. These item-total correlations are subsequently correlated with the degrees to which items show group differences in performance, i.e., the (standardized) group differences between items' p-values or proportions correct.

Figure 1 and Table 1 illustrate MCV at the item level in a scenario involving three groups that have equal variance on the latent trait, but different mean levels of ability. The test in Figure 1 is composed of five items that follow a so-called Guttman scale, which is the most basic type of item-response model in which scores on each item depend solely (and deterministically) on the location of the latent ability scale. Figure 1 also depicts the normally distributed theta values for the three groups. For instance, for Item A, all those who have a latent ability exceeding the difficulty parameter of that item (namely -1) should answer the item correctly, whereas all those with lower ability levels should answer it incorrectly. The difficulty parameters of the remaining items are

0, 0.5, 1, and 2 for Items B through E, respectively. If the mean ability level of a group is equal to the difficulty parameter of the item, 50% of test takers in that group answer that item correctly (given normal latent ability distributions). This applies to item B for the low-scoring group and item D for the middle-scoring group. These items cut the normal theta distribution precisely at their respective means. Because the information value (or its ability to discriminate between ability levels) of an item is at its maximum around p = .50, Item B shows the highest item-total correlation in the low-scoring group. This item, however, is too easy for the high-scoring group (with the p-value being close to 1), and hence does not distinguish well between ability levels *within* that high-scoring group, leading to a very low item-total correlation of .04 for that item in that group. In general, the farther removed the item difficulty parameter is from the latent ability mean of a given group (and hence the nearer the p-value is to either 0 or 1), the lower will the item-total correlation become. This is caused by the fact that item variances are a function of the p-value and at their maximum when p = .50. It is important to note that all five items have the same difficulty *parameter* in the IRT sense within the three groups, and hence that there is full measurement invariance (no DIF) between groups. All I have shown here is that the item-total correlation is necessarily different for groups differing in mean latent ability. This result is very general and also applies to other IRT models, like the Rasch model or extensions thereof. Wicherts and Johnson (2009) showed similar graphs and results based on the Rasch model.

The differences in CTT item statistics between groups can be large; it is quite possible for a vector of item-total correlations in one group to correlate negatively with the vector of item-total correlations in another group. For instance, the item-total vectors correlate negatively at -.72 between the low and high scoring groups in the example of Figure 1. Generally, larger group differences in latent ability yield more

diverging item-total correlations.

The group-specificity of item-total correlations appears not to be appreciated by those who apply MCV to the item level. For instance, te Nijenhuis, Bakhiet, et al. (2016) referred to Jensen's requirement that vectors of *g* loadings need to be sufficiently similar across groups to apply MCV (as is sensible for MCV applications in the linear factor model with subtest scores) and subsequently applied Principal Components Analysis (PCA) to compare the different vectors of item-total correlations across numerous samples. Their very use of PCA in studying group differences in vectors of item-total correlations indicates that systematic differences in these supposed *g* loadings exist across samples differing in ability. Te Nijenhuis et al. suggested that the differences were due to sampling errors, while in fact the vectors are necessarily different at the population level (even if there is no sampling error).

Table 2 provides correlations between the vectors of 60 item-total correlations from six groups that were used in a recent MCV study by te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al. (2015) to study Spearman's hypothesis on the SPM.[2] The group means on the SPM are also given and the groups are ordered in their mean SPM performance. As can be seen, the vectors of item-total correlations become less similar and even show substantially negative correlations as the group mean scores become more dissimilar. This group-specificity leads to ambiguous results in which Jensen Effects can appear or disappear depending on which vector one uses in MCV. Te Nijenhuis and colleagues often used vectors of item-total correlations from other groups than the ones being compared in their papers. For instance, in their study of adult

[2] The data were shared by Drs. te Nijenhuis and Al-Shahomee. Note that not all groups in the analyses presented in that paper were included here because of the unavailability of the relevant item-total correlations (the authors pooled some of the groups in their analysis to obtain overall proportions correct).

samples, te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev, et al. (2015) based their conclusions regarding Spearman's hypothesis on an analysis using the item-total correlations from a study with the SPM in Estonia among 12-18 year olds (Lynn, Allik, & Irwing, 2004). Te Nijenhuis et al.'s failure to present the relevant vectors of *g* loadings obscured to readers that their crucial test of Spearman's hypothesis depended crucially on their choice of item-total correlations.

5. Item-total correlations are not *g* loadings

The fact that item statistics from CTT like the item-total correlation are unequal across groups differing in ability renders a generic term like "*g* loading of the item" meaningless; an item that has a maximum item-total correlation in one group may be the item with the lowest item-total correlation in another group. This is not because the item taps *g* differently in different groups (which would entail DIF), because *g* is defined *as a latent dimension*, but rather because the CTT item statistics lack the invariance property due to being computed as a within-group correlation based on sum scores.

Seeing the sum score on the Raven's as the score on the targeted construct *g* reflects an incorrect interpretation of the true score which is common in CTT applications (Borsboom & Mellenbergh, 2002). But even the use of an IRT model to study group differences does not guarantee that the *theta* underlying Raven's test performance reflects *g* and nothing else, which is why measurement invariance is a necessary but not a sufficient condition for Spearman's hypothesis. A core problem with the application of MCV to item scores is that the latent ability that underlies the single test may not be *g*, but rather (also) reflect other ability factor(s) (e.g., working memory capacity, fluid reasoning, spatial abilities, etc.). Spearman's notion of the indifference of the indicator applies only to the *g* factor extracted from a *battery* of cognitive tests, but not to single subtests (Jensen, 1992). In Spearman's theory, subtests measured both *g*

and the subtest-specific ability *s*, and group differences in *s* can occur for a host of reasons (Wicherts & Dolan, 2010). For instance, a vocabulary test may show ethnic group differences because of group difference in *g* and/or in narrow vocabulary ability.

Although Raven's Progressive Matrices tests are widely seen as good *indicators* of *g* in western samples (Jensen, 1998), it is insufficiently clear whether this applies also to non-western samples that feature in the dozen of studies using MCV on the SPM. For instance, the results of 10 factor analytic studies with the Raven's tests in samples of sub-Saharan African test-takers (Wicherts, Dolan, Carlson, & van der Maas, 2010) suggested that they were relatively weak and not always factorially pure indicators of *g.* Even in large cognitive test batteries in western samples, the Raven's test certainly not always displayed the highest *g* loading (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004). More importantly, the Raven's tests have been shown to yield the largest Flynn Effects (Flynn, 2007), which is considered by many *not* to be on *g* (te Nijenhuis, 2013). This implies that group differences on the Raven's tests are not always due to *g* (cf. Fox & Mitchum, 2012). If indeed the Flynn effect creates non-*g* differences on the Raven's tests it is not sensible to equate its *theta* or its sum score simply with *g* in studying group differences.

It is an empirical issue whether or not the trait underlying Raven's test performance equals (or accurately reflects) *g* in a given application of MCV. To study this rigorously, one needs to consider the factorial relations between the Raven's score and a *g* factor based on a sufficiently large battery of cognitive tests. Below, I present a study of whether MCV at the item level can yield Jensen Effects when the trait in the second group is clearly not *g*.

6. Relations between vectors are complexly non-linear

MCV with dichotomous items entails correlating the different vectors like those in Table

1. For instance, for the SPM, the 60 item-total correlations are correlated with either the raw difference in p-values between groups, or a standardized version of that group difference such as the phi coefficient. The phi coefficient is the product-moment correlation coefficient between two dichotomous variables in a two-by-two table, which in this case is simply the correlation between item score (correct vs incorrect) and the group indicator. When groups are of equal size, the phi coefficient reaches its maximum value when the proportion correct in the combined sample equals .50. MCV correlations depend on which of the two possible item-total correlation vectors is correlated with the vector of group differences. With the three-group scenario in Figure 1, the MCV correlations varied from .29 to .85 even though it was made to conform perfectly to Spearman's hypothesis. The MCV correlations differ for the different groups in Figure 1 depending on which item-total vector one uses. This divergence of results is due not to any form of DIF or sampling error, but rather to the complex relations between the vectors that emerge in dichotomous data.

The relation between the vectors are complexly non-linear even in the (hypothetical) scenario wherein the items in the scale have ideal psychometric properties (as in the Guttman scale) and Spearman's hypothesis is true (i.e., all items measure only *g* without any DIF). Figure 2 shows results from different two-group scenarios that again conform to Spearman's hypothesis. Here we see various lasso-shaped bivariate relations between phi coefficients and item-total correlations that here solely depend on how the two group differ on the latent trait. The plots are based on the following conditions: in both groups one normally distributed trait underlies test performance and the 60 items in the scale follow a Guttman scale, where the locations (difficulty parameters) of the items are evenly spaced from -5 to 5. This roughly reflects the item difficulties (based on Rasch analyses) in the Raven's Standard Progressive

Matrices or SPM (Raven, 2000). Although the deterministic Guttman model underlying Figure 2 is not expected to fit to actual data of the SPM, stochastic expansions like the Rasch model will yield highly similar patterns (Wicherts & Johnson, 2009). The plots in Figure 2 are based on exact population values and were computed with an Excel file that can be found as an online supplementary file. Readers can use the file to get a sense of the lasso-shaped relation between the vectors that is expected to occur given particular group differences in the latent trait distribution (and/or with alternative values of the item location parameters). For instance, the upper left panel gives the relation between the vectors in MCV when the mean ability is -3 in the low-scoring group and 2 in the high-scoring group (and both groups have a SD of ability of 1.5). Here, the MCV correlation equals .16 if one were to use the item-total correlations of the low-scoring group. Such would probably not be seen as confirmation of Spearman's hypothesis even though it is actually true in all hypothetical scenarios in Figure 2. Note that the MCV correlation would be .39 in the same scenario if one were to use the item-total correlation vector from the high-scoring group, thereby again highlighting that item-total correlations differ across groups differing in latent ability (as we have already seen empirically in Table 2).

The panels in Figure 2 depict different scenarios in which the groups differ in mean and/or in variance of the latent trait. The shapes of the relations between the vectors under Spearman's hypothesis (full measurement invariance) on this "ideal scale" are quite diverse. The two relevant MCV correlations are given in each panel and vary from r = .05 to r = .97. Although in one case the relation approaches linearity, this just happens to be a scenario in which the lasso was flat due to that particular constellation of Ms and SDs of latent ability in the two groups. In all scenarios, the size of the MCV correlation itself depends on the choice of the vector of item-total

correlations, of which one has two per comparison. The arbitrary choice between the two item-total correlation vectors creates ambiguity. In MCV applications, it is common to focus on the results based on the item-total correlations of the higher-scoring samples (some applications involved item-total correlations from a third group, as discussed in Section 7). Although Figure 2 is based on phi coefficients, the use of unstandardized differences in p-values provides highly similar results.

In actual data, item scores are subject to unsystematic and systematic sources of variation (randomness in responding, guessing, etc.), and all statistics will be subject to sampling variation. Figure 3 depicts one of the lassos from Figure 2 with 95% sampling distributions based on the standard SEs for each of the 60 phi coefficients (horizontal lines) and each of the 60 actual item-total correlations (vertical lines) for when $N = 200$ (100 participants per group). This plot highlights the sampling variability that can be expected with a perfect scale like this in a sample of this size. In real data, the relation between trait and items scores is stochastic, and so results will become noisier. Applications of MCV (and even the very definition of Jensen Effects according to Rushton, 1998) are often based on null hypothesis significance tests with the number of items as N (Rushton, 2002; Rushton et al., 2002) or simply on MCV correlations being around .30 or higher (te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al., 2015; te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev, et al., 2015). With 60 items, as in Raven's SPM, all MCV correlations lager than .255 are significant at Alpha = .05. Such standard tests are statistically incorrect in this context, because they do not take into account the nonlinearity and the actual sources of sampling error. Because each item has its own proportion correct and its own item-total correlation, the SEs of these items statistics are different, thereby violating the "identically distributed" assumption underlying standard significance tests.

The main message from Figure 2 and the Excel file that readers can use to study other scenarios is that vast differences in MCV correlations can occur even if all items measure a single cognitive ability in an invariant manner across groups. This variation in MCV correlations is in no way due to Spearman's hypothesis being any less "true" in any of the scenarios, since it is true for all scenarios used as input (again under the assumption that theta equals *g*). The variation in MCV correlations also has relatively little bearing on issues like sample sizes, the reliability of the item-total vectors, or the reliability of the group differences in item performance as te Nijenhuis, Bakhiet, et al. (2016) suggested recently. Rather, the variation in MCV correlations is due to the complex non-linearity that is caused by the restricted range of the relevant CTT item statistics (that go from -1 to 1) and their inherent group-specificity. That MCV correlations based on item scores can be very low (negative even) in the population even if Spearman's hypothesis were true highlights that MCV lacks sensitivity: Even if Spearman were right and the test measures *g* and nothing else, MCV correlations close to 1 will only occur in rare cases. This complexity of the relations between the vectors causes MCV results to be highly unstable and very difficult to interpret.

7. Empirical results showing nonsensical Jensen Effects

7.1 Introduction

The previous section highlighted that MCV applied to item level data often lacks sensitivity, but how about MCV's specificity, i.e., its capability to detect DIF or item bias when Spearman's hypothesis is incorrect? There are several reasons to expect that MCV at the item level will not be able to detect DIF even if it is severe. For starters, one can envision many scenarios in which group differences in item parameters (i.e., DIF) will not lower the MCV correlation. For instance, image we would add 10 DIF items that show a large artificial group difference and function well in the high-scoring group (as

evidenced by a high item-total correlation in that group) to the scenario depicted in the upper-left panel of Figure 2. Because many applications of MCV do not consider the item-total correlations in the low-scoring group (some even use item-total correlations for an entirely different group than the two being compared), the user might not even notice that these DIF items show near zero item-total correlations in the low-scoring group. Nonetheless, such DIF items could heighten the MCV correlation (based on the item-total correlations of the high-scoring group or another group) considerably above the value that we would expect on the basis of no DIF (which was only .39). Thus, there is no reason to expect item bias or DIF against the lower-scoring group to necessarily lower the MCV correlation based on the item-total correlation in the high-scoring group. Because under Spearman's hypothesis (no DIF) the MCV correlation already can assume a wide range of values (see Figure 2), it will be hard to determine whether any given MCV results is due to the particular shape of the relation or to DIF.

Still there exist a statistical artifact that renders MCV correlations to be positive rather than negative in most applications, regardless of whether the items display DIF. The artifact, which was also discussed by Wicherts and Johnson (2009), is caused by the fact that the two main ingredients of MCV, namely item-total correlations and (raw or standardized) group differences in p-values are both sensitive to the p-value, i.e., the proportion correct in the sample. Consequently, these ingredients will tend to correlate positively as long as the items in the two groups are ordered similarly in terms of p-value, and the group differences are not too large.

The item-test correlation depends on the item's standard deviation, which is a direct function of p (namely: $\sqrt{p(1-p)}$). The maximum item-total correlation occurs when the p-value within a group equals .50, simply because it is at its maximum when the SD is at its maximum around .50. These maximum values as a function of the p-value

in the group are displayed in Figure 4 next to the maximum values of the other ingredient of MCV, namely the group differences in item performance (based either on raw mean group difference in p-values or some standardization thereof like the phi coefficient). When groups are of equal size, the value of phi is at its maximum when the p-value in the combined sample equals .50. The relation between the maximum (unstandardized) group difference in p-values and the combined-sample p-value is very similar, and again shows the highest maximum around p = .50 in the combined sample.

Because both vectors in MCV are intimately related to the p-value in the groups, one would expect a positive correlation between them (and perhaps generating a Jensen Effect) when (1) the item-total correlations are positive (as implied in the reliability of the scale, itself being based on average item-total correlations; Kuder & Richardson, 1937) and (2) the item p-values are sufficiently linearly correlated across the two groups. The latter requires that items in the two groups are ordered similarly in terms of difficulty, that the group differences in item performance are not too large, and that there the groups are similar in trait variance. Although not always reported in applications with the Raven's tests, the vectors of p-values across groups typically correlate highly (above r=.7). Any reasonable scale with sufficient spread in item difficulties could provide such a pattern as long as its items are ordered similarly in difficulty in both groups, regardless of which trait is measured. This formal result leads to a risky prediction: Can we obtain Jensen Effects with MCV when the items administered to the two groups are different and measure entirely different traits?

7.2 Methods

The goal of this study was to see whether MCV at the item level can yield Jensen Effects in scenarios in which the items in the first group are based on the SPM and the items in the second group are based on a scale that is composed of items that measure a

trait that has no bearing on *g*, yet has a similar ordering of the items in terms of difficulty. To this end, I used the SPM scores of the samples used in earlier MCV studies (see Tables 2 & 3) and compared those to two samples in which I replaced the 60 SPM items with 60 entirely different non-cognitive items. My reason for using two entirely different scales in applying MCV is not that anyone would actually use MCV in such a scenario, but rather to study whether MCV is capable of rejecting Spearman's hypothesis in cases where the violation of measurement invariance is arguably most severe because the items are measuring entirely different traits in the two groups. Thus, this is an empirical test of MCV's specificity: if it cannot detect a failure of invariance in this extreme scenario, it is unlikely to detect less strong violations of invariance in other instances wherein Spearman's hypothesis is untrue.

In order to obtain scales that were comparable to the design of the SPM, I looked for data that were collected as part of the freshmen-testing program at the Psychology Department of the University of Amsterdam. In the years from which I have data (I coordinated the program as a student around the turn of the millennium), there were two questionnaires with a sufficient number of items that I could use to mimic the 60-item SPM. The first test was composed of items of the Dutch translations of the 40-item State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983) and the 20-item State-Trait Anger Scale (STAS; Spielberger, Jacobs, Russell, & Crane, 1983). For this test, I had complete data from 528 test-takers with an average age of 21 (henceforth the STAI/STAS sample).

The second useful test from the Amsterdam testing program was the 'Vijf PersoonlijkheidsFactoren Test' or 5PFT (Elshout & Akkerman, 1975), which consists of 70 items and is one of the first personality tests that was specifically developed to measure the Big Five Personality Factors Extraversion, Agreeableness,

Conscientiousness, Neuroticism, and Openness to Experience. For the 5PFT (henceforth the 5PFT sample), I had data from 6676 psychology freshmen (incidentally also including yours truly), which were used in an earlier publication (Smits, Dolan, Vorst, Wicherts, & Timmerman, 2011) and are now freely available for anyone to (re)use (Smits, Dolan, Vorst, Wicherts, & Timmerman, 2013).[3]

Because answers on the 5PFT and the STAI/STAS were given on seven-point and four-point Likert scales, respectively, I needed to recode these scores to obtain dichotomous scores for MCV computations involving the SPM samples as comparison groups. To this end, I recoded the first option on the 60 STAI/STAS items to zero and options 2-4 of this scale to one. With the 5PFT, I recoded options 1 and 2 to zero and the remaining five options to one. I used the first 60 (out of 70) items in the 5PFT, but because of the original ordering of items in the 5PFT, these 60 items were evenly distributed across the five personality factors. Contraindicative items in the original scales were already reverse recoded before my analyses. Next, I reversely coded all 12 Neuroticism items from the 5PFT to obtain a set of 60 5PFT items that showed positive item-total correlations. Arguably, the trait measured by these 60 5PFT items can be considered a measure of the general factor of personality. The trait underlying the 60 STAI/STAS items is more ambiguous and might measure a general factor of being anxious and angry both as a trait and a state. The real substance of the two traits underlying the STAI/STAS and 5PFT is not of interest, as long as the newly formed scales show some internal consistency and the items are ordered in terms of difficulty. Nonetheless, I did verify the correlation of the newly formed scales with scores on the Raven's Advanced Progressive Matrices that (part of) both samples also completed

[3] Additional openly available data of the 5PFT (Wicherts & Bakker, 2012) can also be used to correlate the 5PFT with an actual *g* factor derived from seven cognitive tests.

during the testing program (under a 20-minute time limit). Appendix B reports sensitivity analyses for both the STAI/STAS and 5PFT samples that addressed whether results were sensitive when using another way of dichotomizing the Likert scales.

Because the items in the SPM are ordered in five sets of a dozen items that are each ordered in difficulty, I ordered the STAI/STAS and 5PFT items accordingly on the basis of independent samples. To get those independent samples, I used the total samples (N = 528 for the STAI/STAS and N = 6776 for the 5PFT), ordered cases on the basis of a random number between 0 and 1, and selected the first half of both the 5PFT sample and the STAI/STAS sample. These independent samples were composed of 276 and 3346 respondents for the STAI/STAS and 5PFT, respectively. I used these independent samples to order the items in both new scales in terms of difficulty. Specifically, I first determined the dichotomized item's p-values for both scales in these independent samples and subsequently rank-ordered 12 items in terms of p-values in each of five sets of 12 items. I then rank-ordered these five sets according to the overall sum score, with easiest items in the first set of 12 items and the most difficult items in the last set of 12 items. In this way, the rank-order of the 60 items in the newly formed scales mimicked the rank-order of the items in terms of difficulty in the SPM.

Subsequently, I used the remaining respondents to conduct the MCV analyses. The final STAI/STAS sample included 252 respondents and the final 5PFT sample included 3330 respondents. Because the item orderings were based on separate (independent) samples, they can be considered a *design feature* of the two new 60-item scales measuring anxiety and anger (STAI/STAS) and personality (5PFT). The p-values and the item total correlations from the analysis samples are given in Tables A1 and A2 in Appendix A. All the raw data is also uploaded to the OSF page accompanying the current article. Notwithstanding the heterogeneity of their items, the two new scales

yielded a Cronbach's Alphas (KR-20 values) of .94 in the STAI/STAS sample and .76 in the 5PFT sample. This reminds us that internal consistency is uninformative concerning the dimensionality of the scale.

Origins, sample sizes, age ranges, and scale reliabilities for the six SPM samples are given in Table 3. Most of these samples also featured in a previous MCV study by te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al. (2015) (see also Table 2). Specifically, I retrieved item-level data from Rushton et al. (2002) and Rushton, Cvorovic, et al. (2007) and also used raw data from a sample of Libyan secondary school children shared kindly with me by dr. Al-Shahomee (Al-Shahomee, Lynn, & Abdalla, 2013). In addition, I obtained p-values and item-total correlations from the group of Libyan university students from dr. te Nijenhuis. These data sets enabled me to compare item-level data from Dutch psychology freshmen on the STAI/STAS and the 5PFT with the SPM scores of Indian, White, and African engineering students from South Africa (Rushton et al., 2002), Roma adults (Rushton, Cvorovic, et al., 2007), Libyan 16-year-old secondary school students (Al-Shahomee et al., 2013), and Libyan university students (te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al., 2015).

7.3 Results

Before conducting the MCV analyses, I wanted to ascertain that the two newly formed scales had no bearing on *g*. A total of 244 students in the STAI/STAS sample also took the actual Raven's Advanced Progressive Matrices (APM). Their APM scores (M = 21.2, SD=4.2) did not significantly correlate with their sum score on the 60-item STAI/STAS scale: r = -.09, p = .168). Similarly, APM scores of 505 students (M = 21.4, SD = 4.5) in the total 5PFT sample weakly correlated (r = -.02, p = .652) with the 5PFT sum score. I also correlated the 60 individual dichotomous item scores from the 5PFT with the score on the Raven's APM, which provided correlations that were -.002 on average.

These correlations of individual 5PFT items with the APM themselves correlated at r = .050 (N = 60) with the item-total correlations that were used as "g loadings" in the MCV analysis. The same analysis of the 60 STAI items yielded very low correlations with the Raven's APM (mean r across the items of -.040) and again showed no meaningful correlation between the item-total vector from the STAI/STAS (i.e., the supposed g loadings in the MCV analyses) with the APM sum score (r = .057, N = 60). Assuming that the APM does have a substantial g loading in these samples (which it appears to do; Wicherts & Bakker, 2012), these analyses ensured that the item-total correlations based on the STAI/STAS and the 5PFT cannot be interpreted as g loadings, and that both newly formed scales had no meaningful relation with g.

The mean scores and reliabilities of the scales in the eight groups are given in Table 3. The MCV correlations for the twenty group comparisons are given in Table 4. The MCV correlations based on the item-total correlations of the higher-scoring group are given below the diagonal, and the MCV correlations based on the item-total correlations from the lower-scoring groups are given above the diagonal. The order in which the six groups appear is based on their mean sum scores (see Table 3). Note that most MCV applications focus on the results using the item-total correlations of the higher-scoring group, which led me to focus particularly on the results below the diagonal in Table 4.

te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al. (2015) reported that the average MCV correlation in 11 studies of the items of Raven's Progressive Matrices was r = .30, and so this value can be used as a yardstick to assess the current results. The comparisons of samples that completed the SPM with other samples that completed the SPM shows similar results (not surprising given the overlap in samples) of MCV correlations lying between -.18 (for the White students vs. Libyan 2nd school students)

and .81 (for Roma adults vs. Libyan students). These MCV correlations of SPM-SPM comparisons center around .35. MCV correlations based on the item-total correlations from the higher-scoring groups (given below the diagonal in Table 4) were generally higher than those based on the item-total correlations of the lower-scoring groups (given above the diagonal in Table 4), which might be due to the means in all samples exceeding 30 (readers can assess this when using the Excel file).

The second column of results in Table 4 contains the MCV correlations when comparing the STAI/STAS sample with the six groups that took the SPM. These correlations were around .30: r = .26, .33, .31, .32, .24, and .34. Five of these correlations would have been significant when tested against the standard null hypothesis of $r = 0$, and so they *are* Jensen Effects as defined by Rushton (1998). If we only interpret MCV correlations based on the item-total correlations in the high-scoring groups (given below the diagonal), four of the MCV correlations comparing STAI/STAS with the SPM were significant and larger than .255.

The analyses in Table 4 were based on pairwise deletion of items that were answered correctly or incorrectly by the entire sample at hand, because this impedes the computation of the item-total correlation. However, imputing zero values for item-total correlations that were undefined for that reason yielded highly similar results. In fact, the MCV correlations for the STAI/STAS comparisons were even somewhat higher (r = -.18, .33, .31, .40, .40, and .56) and significant in five cases after imputation. The full MCV correlation results after imputation are given in Appendix A. Furthermore, Appendix B reports another set of sensitivity analyses in which I used another way to dichotomize the Likert items in both the STAI/STAS and the 5PFT samples. These analyses corroborate the results in Table 4.

MCV correlations with the 5PFT sample are given in the sixth row in Table 4 and

replicate these results with another scale that showed higher mean sum scores. Here the relevant MCV correlations (computed using the item-total vectors of the high-scoring group) comparing the 5PFT with the six SPM samples were .54, .50, .52, .31, -.16, .04, and four of these comparisons showed a Jensen Effect as defined by Rushton (1998). Using imputed values for undefined item-total correlations yielded MCV correlations for the 5PFT of .54, .50, .52, .31, -.02, and .01. The latter correlation comparing the 5PFT with South-African White engineering students would have been .51 if I had used the 5PFT "*g* loadings" throughout, which again highlights the ambiguity in testing Spearman's hypothesis.

So substantial Jensen Effects can be easily obtained when the test at hand is completely different in one of the groups. In the current applications of MCV in comparing SPM with SPM scores, the correlations did tend to be somewhat higher than the comparisons between SPM and STAI/STAS and SPM and 5PFT. Nevertheless, the average MCV correlations in comparing the SPM with scores on the current measure of anxiety and anger (r = .23) and the current measure of personality (r = .29) are quite close to the average MCV correlations reported in a recent survey of studies of this kind in the literature (te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al., 2015).

In their own study of the SPM scores of Libyan adults, te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev, et al. (2015) reported correlations that were appreciably higher than the values in Table 4. The data shared with me by dr. te Nijenhuis clarified that this was due to the fact that they did not use the item-total correlations from the high-scoring groups but rather the item-total correlations from another sample. Specifically, they used the item-total correlations in a group of 2,735 12-18 year-old high school students from Estonia (Lynn et al., 2004). This sample was referred to as "the largest available White group", although te Nijenhuis, Al-Shahomee, van den Hoek,

Grigoriev, et al. (2015) did not refer to the relevant article by Lynn and coworkers. The argument for using this vector appears to have been that it was more reliable in lieu of the large sample size on which it is based (notwithstanding the age difference between the adult groups in the comparison and the Estonian sample). Applying the same logic, we could just as well use the item-total correlations from the 5PFT data, which is the largest sample of the current comparisons, with an *N* of 3330.

The use of the item-total correlations from the relatively heterogeneous group of Estonian teenagers in the comparison of adult samples heightened the correlations between the vectors. The reason is that this vector of item-total correlations is based on a wider range of ability and scores that cover much of the score range on the SPM. But any hopes that the use of such a vector of item-total correlations might restore our trust in MCV quickly dissipated when I correlated all group differences with the same item-total correlation vector from Lynn et al. (2004). Results of these analyses are presented below the diagonal in Table 5. I also used the item-total correlations from the largest sample (the 5PFT sample) to re-compute MCV correlations for all group comparisons, and these results are given in the upper triangle of Table 5.

Bar some exceptions, these MCV correlation when using the Lynn et al. (2004) item-total correlations from Estonia and the 5PFT item-total correlations were high. The MCV correlations based on the Estonian "*g* loadings" were between .13 (for a group comparison wherein group differences were tiny) and .81 (for the largest group difference) for the 5PFT data, whereas the MCV correlations for the STAI/STAS were between -.15 and .77. The average MCV correlations using the Lynn et al. (2004) item-total correlations were .34 for analyses comparing the STAI/STAS with SPM samples and .50 for analyses comparing the 5PFT sample with the SPM samples.

Thus, these results showed that one can obtain large Jensen Effects using the

same method as in te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev, et al. (2015) when the two scales are composed of entirely different types of items and measure entirely different traits. The MCV correlation of .62 in the lower half of Table 5 is particularly noteworthy, because this Jensen Effect is on a par with the summary of MCV results in the US as reported by Jensen (1998). Yet this particular MCV correlation compared scores from Dutch university students on a measure of the general factor of personality (5PFT) with scores of another group of Dutch university students on a combined measure of anxiety and anger (STAI/STAS) using the item-total correlations based on a fluid intelligence measure (the SPM) administered to Estonian teenagers. Together with the other Jensen Effects obtained with using the two new non-cognitive scales, this result illustrates that MCV can be entirely insensitive to the fact that the items were completely different across the groups.

8. Conclusions

Wicherts and Johnson (2009) discussed why the method of correlated vectors is deeply flawed when applied to dichotomous items. The core problem is that the method uses CTT item statistics that have been known for a long time to depend on the proportion of test takers who provide correct answers in a given sample. This means that MCV results are dependent on which item-total correlation vector one uses, which introduces ambiguity in the studying Spearman's hypothesis. I argued that MCV applied to item level data does not provide any test of measurement invariance and is flawed because it assumes linearity where linearity clearly does not apply.

The formal arguments that I presented were based on a perfect scale and on population values with known psychometric and statistical properties. As discussed, patterns in real data look much noisier due to the fact that psychometric scales do not function deterministically as in the Guttman scale and because sampling error is always

at play, as can be seen in Figure 3. Nevertheless, if a perfect scale can already provide MCV correlations close to zero even if Spearman's hypothesis were correct and the scale at hand is fully measurement invariant across groups, it is safe to conclude that MCV at the item level lacks sensitivity; even in large samples it will not show a (significant) Jensen Effect in many scenarios. The relevant MCV correlation in scenarios where there is no DIF can vary enormously because of the non-linearity involved (see Figure 2). Thus, MCV has little to no diagnostic value in studying measurement invariance at the item level, which is a core requirement for any comparison of latent variables across groups. A test of Spearman's hypothesis with MCV requires tests or items to be invariant, as even Jensen (1998, p. 374) acknowledged. So MCV cannot be used to test for measurement invariance.

The psychometric problems concerning the use of MCV at the item level are quite severe and go beyond that lack of sensitivity in corroborating Spearman's hypothesis and the hypothesis of measurement invariance that it implies. Namely, just like in applications of MCV at the subtest level, MCV at the item level can provide Jensen effects even if Spearman's hypothesis is not true and massive measurement bias exists. This lack of specificity is not only based on formal psychometric arguments but also on the empirical data showing that clear and nearly consistent Jensen Effects can be obtained when the scale in one of the groups does not measure *g* (or whatever trait the Raven's SPM measures), but rather obscure combinations of anger, anxiety, and personality. Although it could happen that in some applications of MCV at the item level the MCV correlation is larger, even these are difficult to interpret because one could envision scenarios in which biased items can enhance rather than lower that MCV correlation. My current findings do not imply that the method always provides the incorrect conclusion about Spearman's hypothesis, but rather that it often yields both false

negatives (no Jensen Effect despite Spearman being right) and false positives (a Jensen Effect despite a violation of Spearman's hypothesis). Thus, MCV applied to item data has little diagnostic value.

Item statistics' sensitivity to p-values in classical test theory (e.g., see Ferguson, 1941) eventually led to the development of modern IRT (Lord & Novick, 1968). In IRT, item characteristics are defined in relation to the latent ability, thereby enabling meaningful group comparisons as well as rigorous tests of the degree to which items function differentially across groups (Holland & Wainer, 1993). Tests of DIF in IRT are well established and are preferred in the study of group differences.

te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev, et al. (2015) recently used MCV to compare the performance of Libyan adults on the SPM to scores from different samples in South Africa, Spain, Russia and Serbia and concluded that "Spearman's hypothesis was strongly confirmed" (p. 114). Results in Tables 4 and 5 effectively replicated their results after I replaced the SPM with heterogeneous measures of anxiety and anger and the big five personality factors. These newly formed scales had no bearing on *g* yet showed Jensen Effects and large MCV correlations because I ensured that the items were ordered in terms of difficulty as they are in the SPM. In all comparisons of SPM with the SPM the same item ordering applies, and so it could very well be the case that similar levels of DIF have occurred in previous studies that used MCV to compare groups on the SPM. It is high time to redo these studies using rigorous tests for DIF.

Ethnic group differences on the Raven's tests can be large (e.g., Brouwers, van de Vijver, & van Hemert, 2009; Wicherts et al., 2010), and understanding their meaning is important for theory and practice. The current results cast doubt on conclusions drawn

on the basis of Jensen's method of correlated vectors applied to Raven's test items. MCV not only lacks specificity in the sense that it may lead to incorrect conclusions that Spearman was right about group differences (Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004), but also is often insensitive to find support for his hypothesis in item level data if Spearman were right. The best solution for now would be to abandon MCV altogether and use established methods from modern IRT instead when studying group differences in cognitive tests like the Raven's.

References

Al-Shahomee, A. A., & Lynn, R. (2010). IQs of Men and Women and of Arts and Science Students in Libya. *Mankind Quarterly, 51*, 154-158.

Al-Shahomee, A. A., Lynn, R., & Abdalla, S. E.-g. (2013). Dysgenic fertility, intelligence and family size in Libya. *Intelligence, 41*, 67-69. doi: 10.1016/j.intell.2012.11.001

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence, 33*, 431-444. doi: 10.1016/j.intell.2004.12.004

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30*, 505-514. doi: 10.1016/S0160-2896(02)00082-X

Brouwers, S. A., van de Vijver, F. J. R., & van Hemert, D. A. (2009). Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences, 19*, 330-338. doi: 10.1016/j.lindif.2008.10.006

Díaz, A., Sellami, K., Infanzón, E., Lanzón, T., & Lynn, R. (2012). A comparative study of general intelligence in Spanish and Moroccan samples. *The Spanish journal of psychology, 15*, 526-532. doi: 10.5209/rev_SJOP.2012.v15.n2.38863

Dolan, C. V. (1997). A note on Schoenemann's refutation of Spearman's hypothesis. *Multivariate Behavioral Research, 32*, 319-325. doi: 10.1207/s15327906mbr3203_4

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35*, 21-50. doi: 10.1207/S15327906MBR3501_2

Dolan, C. V., & Hamaker, E. L. (2001). Investigating Black-White differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors. In F. Columbus (Ed.),

Advances in psychology research (Vol. 6, pp. 31-59). Huntington, NY: Nova Science Publishers, Inc.

Dolan, C. V., & Lubke, G. H. (2001). Viewing Spearman's hypothesis from the perspective of multigroup PCA: A comment on Schoenemann's criticism. *Intelligence, 29*, 231-245. doi: 10.1016/S0160-2896(00)00054-4

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman's hypothesis: The GAT-B in Holland and the JAT in South Africa. *Intelligence, 32*, 155-173. doi: 10.1016/j.intell.2003.09.001

Elshout, J. J., & Akkerman, A. (1975). *Vijf Persoonlijkheids-factoren Test 5PFT, handleiding [Five personality factors test 5PFT, manual]*. Nijmegen, The Netherlands: Berkhout Nijmegen B.V.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika, 6*, 323-329.

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn Effect*. Cambridge, UK: Cambridge University Press.

Fox, M. C., & Mitchum, A. L. (2012). A knowledge-based theory of rising scores on "culture-free" tests. *Journal of Experimental Psychology: General, 142*, 979-1000. doi: 10.1037/a0030155

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: John Wiley & Sons, Inc.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications.

Jensen, A. R. (1980). *Bias in mental testing*. London: Methuen & Co., Ltd.

Jensen, A. R. (1985). The nature of the Black–White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8*, 193-263.

Jensen, A. R. (1992). Commentary: Vehicles of g. *Psychological Science, 3*, 275-278.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT, US: Praeger Publishers/Greenwood Publishing Group, Inc.

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: consistent results from three test batteries. *Intelligence, 32*, 95-107. doi: 10.1016/S0160-2896(03)00062-X

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company.

Lubke, G. H., Dolan, C. V., & Kelderman, H. (2001). Investigating group differences on cognitive tests using Spearman's hypothesis: An evaluation of Jensen's method. *Multivariate Behavioral Research, 36*, 299-324. doi: 10.1207/S15327906299-324m

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003a). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*, 543-566. doi: 10.1016/S0160-2896(03)00051-5

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003b). Weak measurement invariance with respect to unmeasured variables: An implication

of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology, 56*, 231-248. doi: 10.1348/000711003770480020

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence, 32*, 411-424. doi: 10.1016/j.intell.2004.06.007

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543. doi: 10.1007/BF02294825

Millsap, R. E. (1997). The investigation of Spearman's hypothesis and the failure to understand factor analysis. *Cahiers de Psychologie Cognitive, 16*, 750-757.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48. doi: 10.1006/cogp.1999.0735

Rushton, J. P. (1998). The "Jensen Effect" and the "Spearman-Jensen hypothesis" of Black-White IQ differences. *Intelligence, 26*, 217-225. doi: 10.1016/S0160-2896(99)80004-X

Rushton, J. P. (2002). Jensen effects and African/Coloured/Indian/White differences on Raven's Standard Progressive Matrices in South Africa. *Personality and Individual Differences, 33*, 1279-1284. doi: 10.1016/S0191-8869(02)00012-0

Rushton, J. P., Bons, T. A., Vernon, P. A., & Cvorovic, J. (2007). Genetic and environmental contributions to population group differences on the Raven's Progressive Matrices estimated from twins reared together and apart. *Proceedings of the Royal Society: B, 274*, 1773-1777. doi: 10.1098/rspb.2007.0461

Rushton, J. P., Cvorovic, J., & Bons, T. A. (2007). General mental ability in South Asians: Data from three Roma (Gypsy) communities in Serbia. *Intelligence, 35*, 1-12. doi: 10.1016/j.intell.2006.09.002

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11*, 235-294. doi: 10.1037/1076-8971.11.2.235

Rushton, J. P., & Skuy, M. (2000). Performance on Raven's Matrices by African and White university students in South Africa. *Intelligence, 28*, 251-265. doi: 10.1016/S0160-2896(00)00035-0

Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's Advanced Progressive Matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment, 12*, 220-229. doi: 10.1111/j.0965-075X.2004.00276.x

Rushton, J. P., Skuy, M., & Fridjhon, P. (2002). Jensen effects among African, Indian and White engineering students in South Africa on Raven's standard progressive matrices. *Intelligence, 30*, 409-423. doi: 10.1016/S0160-2896(02)00093-4

Rushton, J. P., Skuy, M., & Fridjhon, P. (2003). Performance on Raven's Advanced Progressive Matrices by African, East Indian, and White engineering students in South Africa. *Intelligence, 31*, 123-137. doi: 10.1016/S0160-2896(02)00140-X

Smits, I. A. M., Dolan, C. V., Vorst, H. C., Wicherts, J. M., & Timmerman, M. E. (2013). Data from 'Cohort Differences in Big Five Personality Factors Over a Period of 25 Years'. *Journal of Open Psychology Data, 1*, e2. doi: 10.5334/jopd.e2

Smits, I. A. M., Dolan, C. V., Vorst, H. C. M., Wicherts, J. M., & Timmerman, M. E. (2011). Cohort differences in big five personality factors over a period of 25 years. *Journal of Personality and Social Psychology*. doi: 10.1037/a0022874

Spielberger, C. D., Gorsuch, R. L., Lushene, P. R., Vagg, P. R., & Jacobs, G. (1983). *Manual for the State-Trait Anxiety Inventory Form Y*. Palo Alto, CA: Consulting Psychologists Press, Inc.

Spielberger, C. D., Jacobs, G., Russell, S., & Crane, R. S. (1983). Assessment of anger: The state-trait anger scale. *Advances in personality assessment, 2*, 159-187.

te Nijenhuis, J. (2013). The Flynn effect, group differences, and g loadings. *Personality and individual differences, 55*, 224-228. doi: 10.1016/j.paid.2011.12.023

te Nijenhuis, J., Al-Shahomee, A. A., van den Hoek, M., Allik, J., Grigoriev, A., & Dragt, J. (2015). Spearman's hypothesis tested comparing Libyan secondary school children with various other groups of secondary school children on the items of the Standard Progressive Matrices. *Intelligence, 50*, 118-124. doi: 10.1016/j.intell.2015.03.002

te Nijenhuis, J., Al-Shahomee, A. A., van den Hoek, M., Grigoriev, A., & Repko, J. (2015). Spearman's hypothesis tested comparing Libyan adults with various other groups of adults on the items of the Standard Progressive Matrices. *Intelligence, 50*, 114-117. doi: 10.1016/j.intell.2015.03.001

te Nijenhuis, J., Bakhiet, S. F., van den Hoek, M., Repko, J., Allik, J., Žebec, M. S., et al. (2016). Spearman's hypothesis tested comparing Sudanese children and adolescents with various other groups of children and adolescents on the items of the Standard Progressive Matrices. *Intelligence, 56*, 46-57. doi: 10.1016/j.intell.2016.02.010

te Nijenhuis, J., Grigoriev, A., & van den Hoek, M. (2016). Spearman's hypothesis tested in Kazakhstan on the items of the Standard Progressive Matrices Plus. *Personality and Individual Differences, 92*, 191-193. doi: 10.1016/j.paid.2015.12.048

te Nijenhuis, J., Jongeneel-Grimen, B., & Kirkegaard, E. O. (2014). Are Headstart gains on the g factor? A meta-analysis. *Intelligence, 46*, 209-215. doi: 10.1016/j.intell.2014.07.001

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence, 40*, 73-76. doi: 10.1016/j.intell.2012.01.004

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis; An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice, 29*, 39-47. doi: 10.1111/j.1745-3992.2010.00182.x

Wicherts, J. M., Dolan, C. V., Carlson, J. S., & van der Maas, H. L. J. (2010). Raven's tests performance of Africans: Average performance, psychometric properties, and the Flynn Effect. *Learning and Individual Differences, 20*, 135-151. doi: 10.1016/j.lindif.2009.12.001

Wicherts, J. M., & Johnson, W. (2009). Group differences in the heritability of items and test scores. *Proceedings of the Royal Society: B, 276*, 2675-2683. doi: 10.1098/rspb.2009.0238

Table 1

Guttman scale item parameters and classical test theory item statistics for five items in three groups as depicted in Figure 1 showing the sensitivity of CTT statistics to latent ability levels.

			Low		Middle		High		phi coefficients		
		DIFF	p	itc	p	itc	p	itc	L-H	M-H	L-M
Item	A	-1.00	.98	.31	1.00	.02	1.00	.00	.11	.00	.11
	B	0.00	.50	.85	.98	.43	1.00	.04	.58	.11	.54
	C	0.50	.16	.78	.84	.78	1.00	.17	.85	.29	.68
	D	1.00	.02	.43	.50	.85	0.98	.43	.95	.54	.54
	E	2.00	.00	.02	.02	.31	.50	.95	.58	.54	.11

Notes: DIFF: IRT difficulty parameter that is invariant across groups; p: p-value in the given group; itc: item-total correlation in a given group; phi coefficient relate to comparing the groups on that item: LH: low-scoring group vs. high-scoring group; M-H: middle-scoring vs high-scoring group; L-M: low-scoring group vs. middle-scoring group.

Table 2.

Correlations between item-total correlations across six samples that took the SPM and the mean sum score on the SPM in each sample

	Roma	Libyan 1	Libyan2	African	Indian	White	M
Roma	1	.69	.77	.15	-.10	-.30	29.3
Libyan 1	.69	1	.78	.26	-.13	-.27	40.3
Libyan 2	.77	.78	1	.45	.08	-.12	40.6
African	.02	.11	.27	1	.52	.42	49.7
Indian	-.22	-.23	-.07	.26	1	.46	52.5
White	-.45	-.49	-.43	.42	.18	1	56.1

Notes: Samples described in Table 3 and derived from te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al. (2015); correlations below diagonal based on pairwise deletion of undefined item-total correlations (due to p=0 or p=1) (Ns varying from 30 to 60); correlations above diagonal based on vectors in which undefined item-total correlations (due to p=0 or p=1) were set at 0 (N=60).

Table 3.

Origins, references, and descriptive statistics for samples used to test Jensen Effects with MCV

Sample	Origin	N	M	rel.	age
Roma adults	(Rushton, Cvorovic, et al., 2007)	231	29.3	.91	16-66
Dutch University students STAI/STAS	[current paper]	252	31.3	.94	17-25
Libyan 2nd school students	(Al-Shahomee et al., 2013)	592	40.3	.92	16
Libyan university students	(Al-Shahomee & Lynn, 2010)	800	40.6	?	18-21
African university students	(Rushton et al., 2002)	198	49.7	.87	17-23
Dutch University students 5PFT	(Smits et al., 2013)	3330	51.8	.76	18-25
Indian university students	(Rushton et al., 2002)	58	52.5	.82	17-23
White university students	(Rushton et al., 2002)	86	56.1	.61	17-23

Notes: all samples completed the SPM except for the Dutch samples that completed the State Trait Anxiety Inventory/State Trait Anger Scale and the 5PFT personality test; rel: KR-20 internal consistency.

Table 4.

MCV correlations for the 28 group comparisons, including those in the Dutch samples based on the STAI/STAS or the 5PFT instead of the SPM.

	Roma	STAI	Libyan 2nd	Libyan st.	African	5PFT	Indian	White
Roma	1	.26	.58	.63	.44	-.03	.28	.07
STAI	-.18	1	.23	.21	-.01	-.27	-.14	-.31
Libyan 2nd	.60	.33	1	.03	.03	-.38	-.03	-.18
Libyan st.	.80	.30	.08	1	.24	-.21	.13	-.04
African	.55	.32	.56	.62	1	.11	.54	.54
5PFT	.54	.46	.50	.52	.31	1	-.02	.51
Indian	.36	.24	.48	.48	.20	-.16	1	.52
White	.39	.34	.44	.57	.55	.04	.48	1

Notes: See Table 3 for sample origins. Libyan 2nd: Libyan secondary school students; Libyan st.: Libyan university students; STAI: STAI/STAS sample; all samples took SPM expect STAI and 5PFT; MCV correlations below diagonal are based on item-total correlations of the higher-scoring sample, MCV correlations above the diagonal are based on the item-total correlations in the lower-scoring sample.

Table 5.

MCV correlations for the 28 group comparisons, including those in the Dutch samples based on the STAI/STAS or the 5PFT instead of the SPM in which the MCV correlations were based on loadings from Lynn et al. (2004) or the 5PFT sample.

	Roma	STAI	Libyan 2nd	Libyan st.	African	5PFT	Indian	White
Roma	1	.22	.06	.07	.37	.54	.51	.65
STAI	.41	1	-.17	-.18	.23	.46	.40	.60
Libyan 2nd	.30	-.15	1	.00	.46	.50	.58	.68
Libyan st.	.36	-.12	.09	1	.51	.52	.61	.71
African	.77	.50	.79	.84	1	.31	.56	.73
5PFT	.81	.62	.62	.63	.18	1	-.02	.51
Indian	.86	.64	.84	.86	.58	.13	1	.57
White	.92	.77	.83	.85	.67	.60	.48	1

Notes: See Table 3 for sample origins. Libyan 2nd: Libyan secondary school students; Libyan st.: Libyan university students; MCV correlations below diagonal are based on item-total correlations of Estonian high-school students (Lynn et al., 2004), MCV correlations above the diagonal are based on item-total correlations from the largest sample (5PFT).

Appendix A

Additional Tables

Table A1

Item statistics of the STAI and STAS (N=252).

Item	p	ITC	Item	p	ITC	Item	p	ITC
stai26	0.92	0.41	stai18	0.55	0.42	stas8	0.59	0.23
stai35	0.82	0.48	stai17	0.54	0.56	stai38	0.48	0.54
stai27	0.78	0.60	stai14	0.46	0.49	stai40	0.54	0.49
stai29	0.68	0.51	stai13	0.36	0.56	stas1	0.33	0.35
stai36	0.68	0.65	stai8	0.84	0.49	stas2	0.35	0.35
stai32	0.64	0.50	stai10	0.77	0.60	stas5	0.27	0.43
stai34	0.69	0.63	stai11	0.72	0.57	stas6	0.25	0.33
stai31	0.66	0.58	stai5	0.72	0.56	stas3	0.24	0.40
stai30	0.63	0.58	stai1	0.71	0.57	stas10	0.81	0.23
stai33	0.62	0.59	stai2	0.54	0.54	stas9	0.66	0.34
stai28	0.48	0.59	stai4	0.65	0.48	stas17	0.37	0.46
stai25	0.37	0.51	stai3	0.57	0.51	stas20	0.17	0.38
stai19	0.82	0.56	stai12	0.37	0.56	stas13	0.13	0.37
stai16	0.72	0.60	stai7	0.38	0.57	stas15	0.13	0.35
stai20	0.73	0.65	stai6	0.29	0.45	stas14	0.16	0.36
stai23	0.76	0.57	stai9	0.19	0.44	stas11	0.09	0.36
stai15	0.75	0.58	stai37	0.81	0.36	stas12	0.08	0.35
stai21	0.71	0.59	stas4	0.71	0.31	stas18	0.07	0.24
stai22	0.66	0.56	stai39	0.75	0.37	stas16	0.06	0.28
stai24	0.66	0.43	stas7	0.67	0.22	stas19	0.06	0.26

Note: ITC: item-total correlation.

Table A2

Item statistics of the 5PFT (N=3300) and the item labels from the original scale.

Item	p	ITC	Item	p	ITC	Item	p	ITC
p27 A	.99	.16	p15 O	.94	.18	p08 C	.91	.31
p42 A	.99	.18	p10 O	.94	.17	p23 C	.88	.27
p22 A	.97	.19	p30 O	.85	.20	p43 C	.89	.24
p12 A	.97	.20	p20 O	.74	.26	p33 C	.85	.27
p57 A	.97	.21	p11 E	.98	.18	p58 C	.83	.30
p37 A	.97	.15	p26 E	.95	.25	p03 C	.68	.31
p17 A	.94	.17	p31 E	.94	.23	p48 C	.69	.38
p52 A	.94	.25	p46 E	.92	.28	p38 C	.59	.31
p07 A	.92	.17	p51 E	.92	.22	p04 N	.80	.39
p02 A	.93	.18	p01 E	.90	.24	p54 N	.79	.37
p47 A	.92	.28	p41 E	.91	.17	p29 N	.76	.41
p32 A	.93	.12	p16 E	.89	.31	p59 N	.75	.35
p60 O	.98	.07	p21 E	.87	.29	p09 N	.76	.37
p25 O	.97	.14	p36 E	.84	.29	p24 N	.72	.34
p45 O	.96	.20	p06 E	.82	.19	p49 N	.70	.38
p55 O	.96	.14	p56 E	.79	.28	p39 N	.65	.39
p40 O	.96	.15	p18 C	.96	.22	p34 N	.66	.34
p35 O	.95	.16	p53 C	.95	.25	p14 N	.64	.38
p50 O	.95	.21	p28 C	.93	.25	p44 N	.60	.34
p05 O	.95	.19	p13 C	.92	.22	p19 N	.52	.35

Note: ITC: item-total correlation; A: Agreeableness item; O: Openness to experience item; E: Extraversion items; C: Conscientiousness item; N: Neuroticism item.

Table A3.

Results of Table 4 in which undefined item-total correlations (because of p=1 or p=0) were imputed with the value 0.

	Roma	STAI	Libyan 2nd	Libyan st.	African	5PFT	Indian	White
Roma	1	.26	.58	.63	.44	-.03	.28	.07
STAI	-.18	1	.23	.21	-.01	-.27	-.14	-.31
Libyan 2nd	.60	.33	1	.03	.03	-.38	-.03	-.18
Libyan st.	.81	.31	.07	1	.29	-.21	.19	.04
African	**.65**	**.41**	**.61**	**.67**	1	**.14**	**.55**	**.56**
5PFT	.54	.46	.50	.52	.31	1	**-.25**	.51
Indian	**.51**	**.30**	**.62**	**.59**	**.26**	**-.02**	1	**.67**
White	**.53**	**.51**	**.62**	**.68**	**.66**	**.01**	**.55**	1

Note: MCV correlations in bold changed because of the use of imputation of undefined item-total correlations.

Appendix B

Sensitivity analysis concerning dichotomizing

Table B1 reports the results of another sensitivity analysis in which I used another way of dichotomizing the items of the STAI/STAS and the 5PFT. Specifically, in this analysis, for the STAI/STAS, I recoded Likert scale values of 1 and 2 to become 1 and Likert values of 3 and 4 to become 0. This yielded a scaled mean score of 50.85 (SD = 8.14) and a Cronbach's alpha of .908 for the newly formed STAI/STAS scale. Moreover, for the 5PFT items, this time I recoded the Likert values 1-4 to the value 0 and the remaining Likert values of 5-7 to 1. This new scale showed a mean of 30.72 (SD = 6.95) and an Alpha reliability of .741. Results in Table B1 clearly replicate the results given in Table 4 in the main text, albeit with the role of the STAI/STAS and 5PFT samples reversed; this time the new scale scores of the 5PFT are relatively low and those from the newly formed STAI/STAS scale were relatively high. As can be seen this sensitivity analysis corroborates the main results.

Table B1

Sensitivity analysis showing MCV correlations for the 28 group comparisons, including those in the Dutch samples based on the STAI/STAS or the 5PFT instead of the SPM.

	Roma	5PFT	Libyan 2nd	Libyan st.	African	STAI	Indian	White
Roma	1	-.14	.58	.63	.44	-.13	.28	.07
5PFT	-.24	1	.36	.32	.12	-.49	-.05	-.38
Libyan 2nd	.60	.64	1	.03	.03	-.40	-.03	-.18
Libyan st.	.80	.53	.08	1	.24	-.34	.13	-.04
African	.55	.27	.56	.62	1	.29	.54	.54
STAI	.59	.45	.46	.48	.26	1	-.13	.15
Indian	.36	.08	.48	.48	.20	-.29	1	.52
White	.39	.19	.44	.57	.55	-.28	.48	1

Notes:. Libyan 2nd: Libyan secondary school students; Libyan st.: Libyan university students; STAI: STAI/STAS sample; all samples took SPM expect STAI and 5PFT; MCV correlations below diagonal are based on item-total correlations of the higher-scoring sample, MCV correlations above the diagonal are based on the item-total correlations in the lower-scoring sample.

Figure 1.

Scenario of a test composed of five items that follow a Guttman scale and three groups that differ in latent ability but not in item parameters (no DIF)

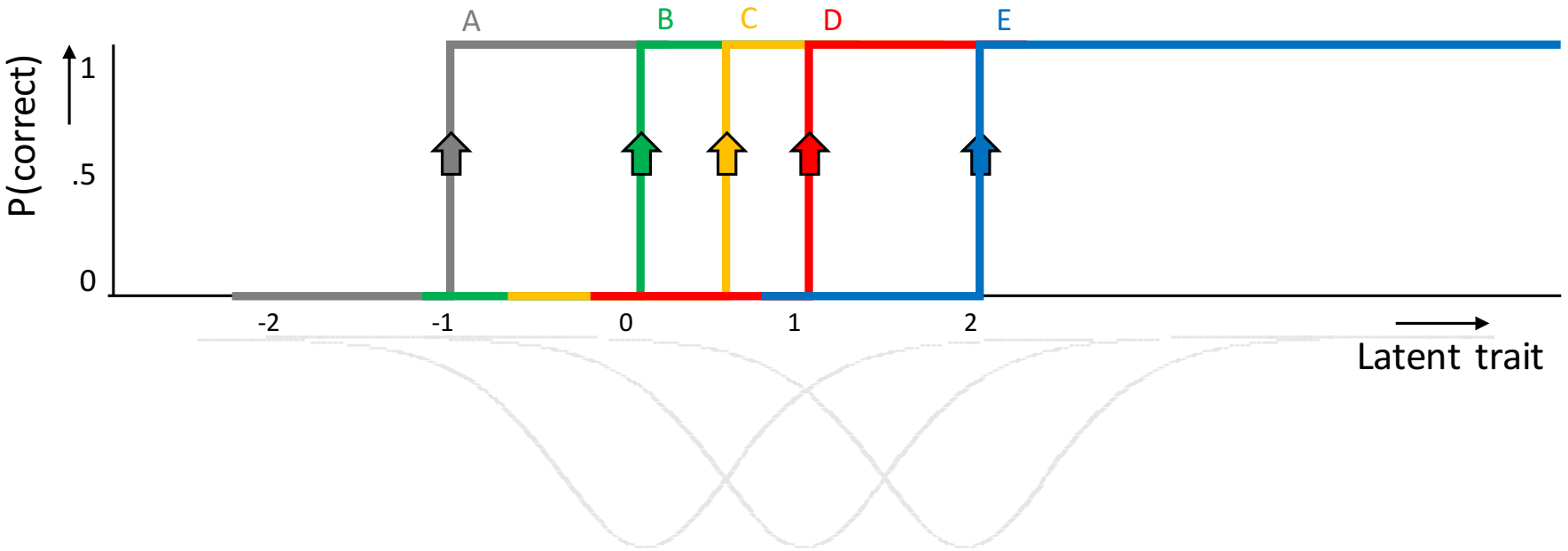

Figure 2.

Different shapes of the relations between the vectors of item-total correlations and phi coefficients used to test Jensen Effects based on a perfect scale of 60 items and two groups differing in latent ability but not in item parameters. Each panel gives another scenario based on particular constellation of Ms and SDs in each of the two groups (given on top) and the resulting MCV correlations (given on the bottom).

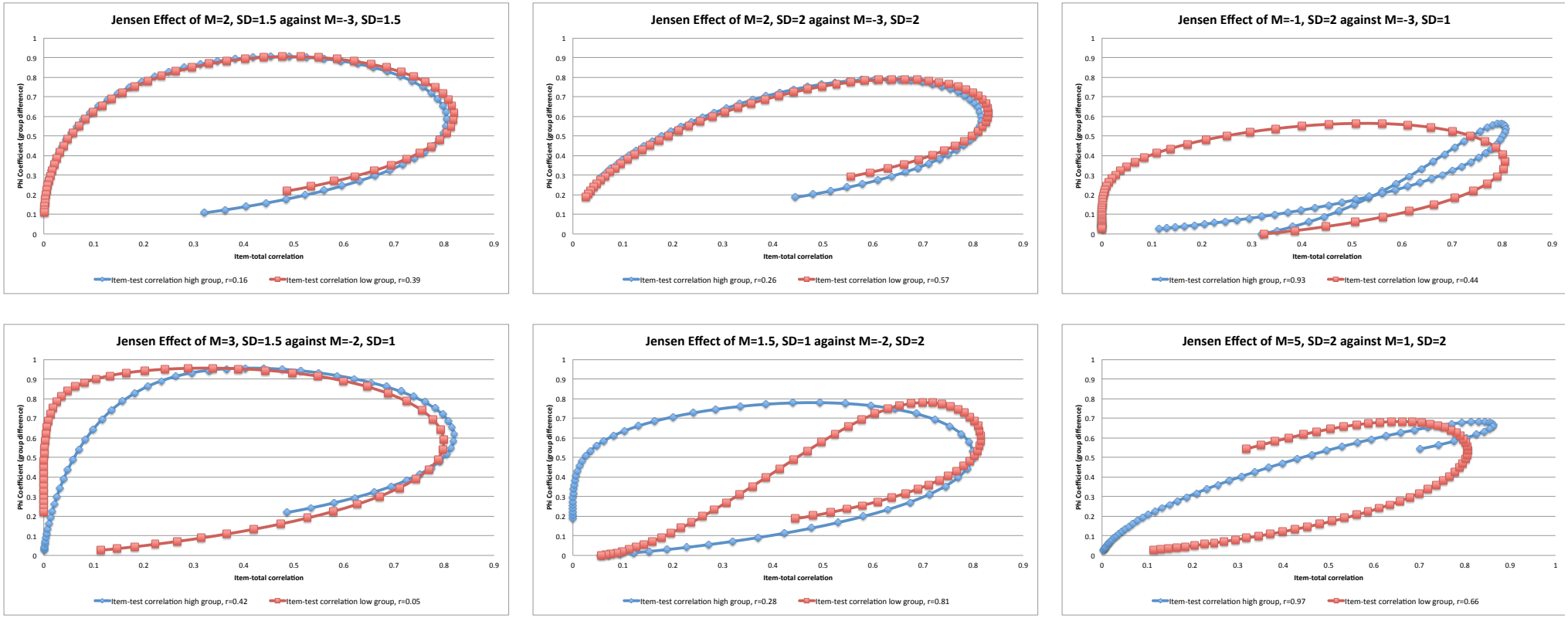

Figure 3.

Relations between the vectors of phi coefficients and item-total correlations and 95% sampling distributions around each item's statistic in samples with N=200

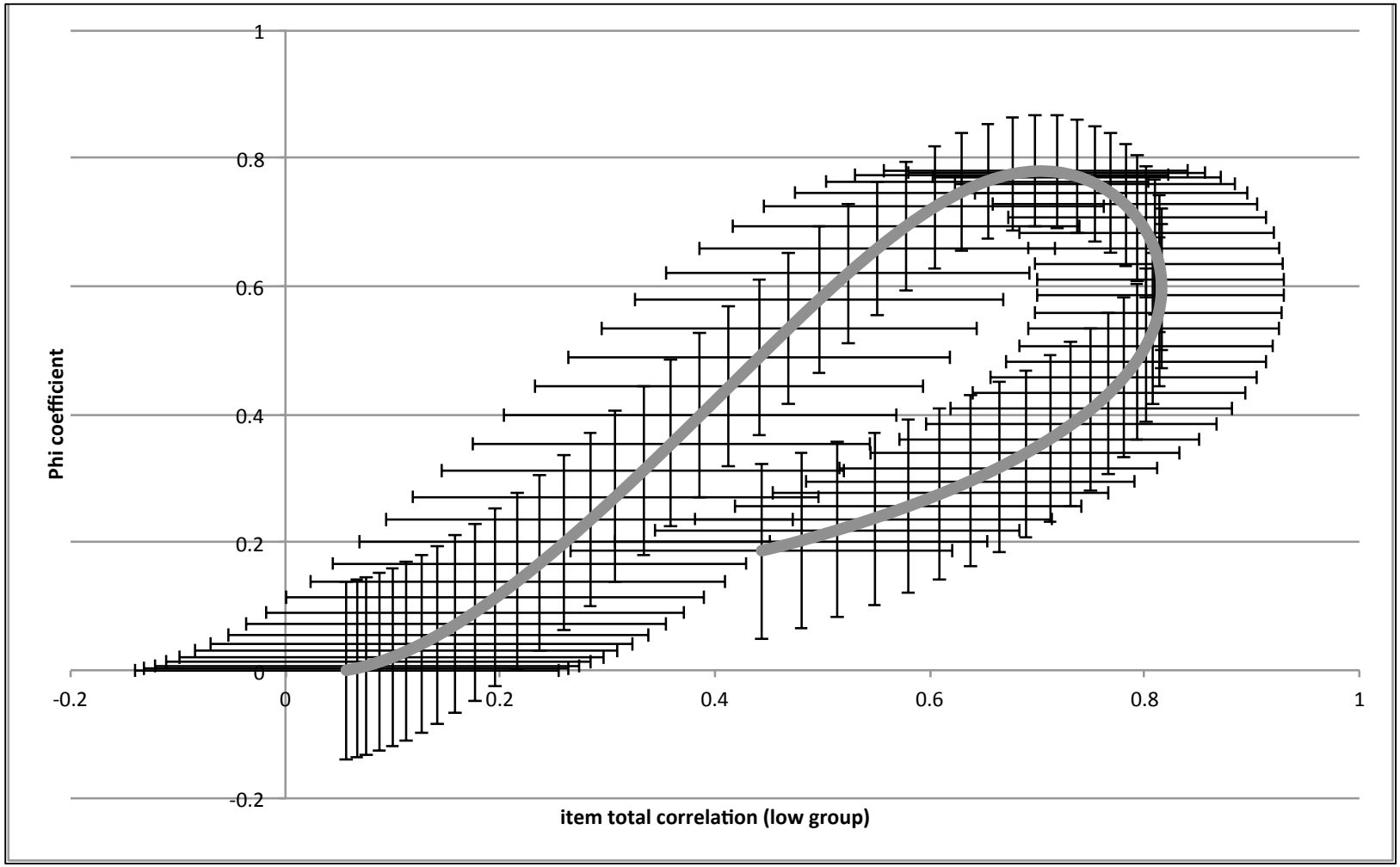

Figure 4.

Maximum values of Phi coefficients, unstandardized group difference in p, and the item-total correlation as a function of the p-value of the items.